 [Translated from English to Vietnamese - www.onlinedoctranslator.com](https://www.onlinedoctranslator.com/en/?utm_source=onlinedoctranslator&utm_medium=docx&utm_campaign=attribution)

Tương tự mã nhị phân thực tế

Phát hiện dựa trên BERT

Học tương tự có thể chuyển nhượng

|  |  |  |  |
| --- | --- | --- | --- |
| Sunwoo Ahn | Seonggwan Ahn | Hyungjoon Koo | Yunheung Paek |
| khoa điện | khoa điện | khoa máy tính | Khoa Điện và |
| và Kỹ thuật máy tính | và Kỹ thuật máy tính | Khoa học và kĩ thuật | Kỹ thuật máy tính |
| Đại học Quốc gia Seoul | Đại học Quốc gia Seoul | Đại học Sungkyunkwan | Đại học Quốc gia Seoul |
| Seoul, Hàn Quốc | Seoul, Hàn Quốc | Suwon, Hàn Quốc | Seoul, Hàn Quốc |

Phát hiện tương tự mã nhị phân (BCSD)

* Vấn đề BCSD



|  |  |  |  |
| --- | --- | --- | --- |
| mã máy | tiệm cận? | mã máy |  |
|  |  |
| đoạn trích A |  | đoạn B |  |

* Nhiều ứng dụng
  + Phát hiện bản sao mã
  + phát hiện phần mềm độc hại
  + Phân loại họ phần mềm độc hại
  + Phát hiện lỗ hổng đã biết
  + Xác minh bản vá mã

thử thách

* Thông tin hữu ích không có sẵn ở dạng nhị phân
  + ví dụ: tên biến, cấu trúc, kiểu, phân cấp lớp, v.v.
* Các tệp nhị phân có ngữ nghĩa giống hệt nhau có thể khác nhau
  + cấu hình trình biên dịch, kiến ​​trúc, obfuscation, v.v.
* vấn đề tạm dừng
  + Không thể quyết định để chứng minh

sự tương đương về chức năng của hai chương trình tùy ý

Công trình hiện có

* Những tiến bộ gần đây sử dụng

mạng lưới thần kinh với kiến ​​trúc Siamese

|  |  |
| --- | --- |
| **Người mẫu** | **Ngành kiến ​​​​trúc** |
|  |  |
| Song Tử | GNN, Xiêm NN |
|  |  |
| Mắt trong | word2vec, LSTM, Xiêm NN |
|  |  |
| Asm2Vec | PV-ĐM |
|  |  |
| Cây cọ | BERT, GNN, Xiêm NN |
|  |  |
| Ngữ nghĩa sâu sắc | Bộ phân loại BERT, Softmax |
|  |  |

Công trình hiện có

* Chức năng khoảng cách/mất mát ảnh hưởng đến mạng Xiêm (Marcelli et al., USENIX '22)

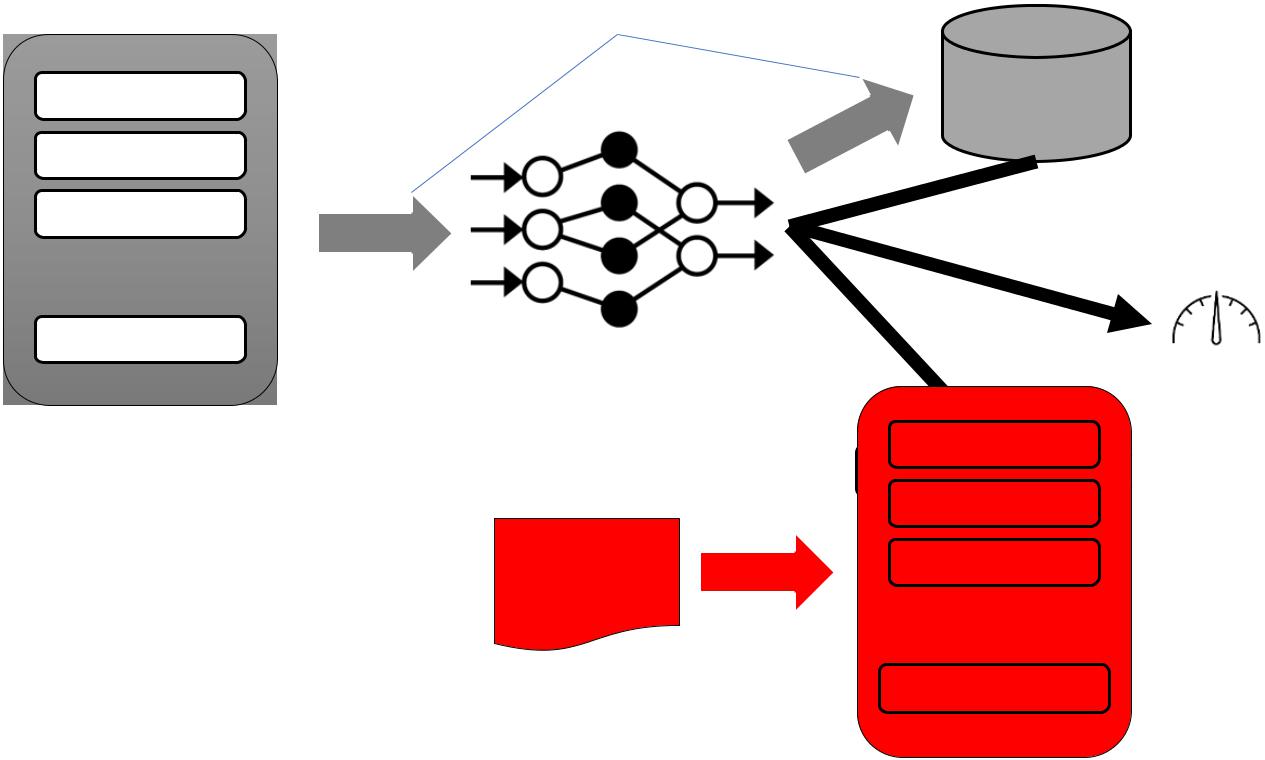
|  |  |  |
| --- | --- | --- |
| **Người mẫu** | **chức năng khoảng cách** | **Mất chức năng** |
|  |  |  |
| Song Tử | khoảng cách cosin | mất tương phản |
|  |  |  |
| Mắt trong | khoảng cách cosin | mất tương phản |
|  |  |  |
| Asm2Vec | khoảng cách cosin | xác suất nhật ký |
|  |  |  |
| Cây cọ | khoảng cách cosin | mất tương phản |
|  |  |  |
| Ngữ nghĩa sâu sắc | Không có | entropy chéo |
|  |  |  |

* giá trị vô hướngđơn giản hóa quá mức

Vấn đề

* Chúng tôi đặt câu hỏi về công việc hiện có trong một kịch bản thực tế

①



chức năng1

chức năng2

chức năng3

|  |  |
| --- | --- |
| … |  |
| hàm\_n | Mô hình BCSD |

Chức năng

nhúng của

sở thích của một người

cơ sở dữ liệu

②

Truy vấn

nhị phân

chức năng1

hàm truy vấn

chức năng2

chức năng3

|  |
| --- |
| … |

tương tự

điểm

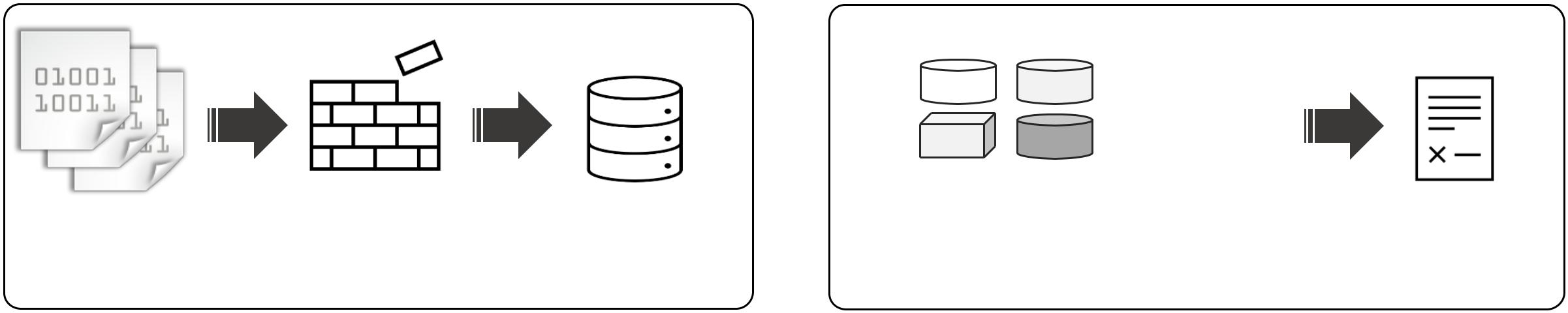
hàm\_m

Phương pháp tiếp cận chính của chúng tôi

* Mục tiêu: cải thiện hiệu suất cho tập dữ liệu chưa thấy
* Học tương tự có thể chuyển giao (dựa trên BERT)
  + Học một mối quan hệ hướng dẫn btw với đào tạo trước
  + Liên tục thể hiện hiệu suất tốt trên ngôn ngữ lắp ráp
* Phát hiện tương tự tốt hơn: học một vectơ khoảng cách có trọng số với entropy chéo nhị phân
* Khoảng cách có trọng sốcác mối quan hệ được biểu diễn trong một vectơ

BinShot

tiền xử lý tinh chỉnh



**Tương tự**

**không giống nhau**

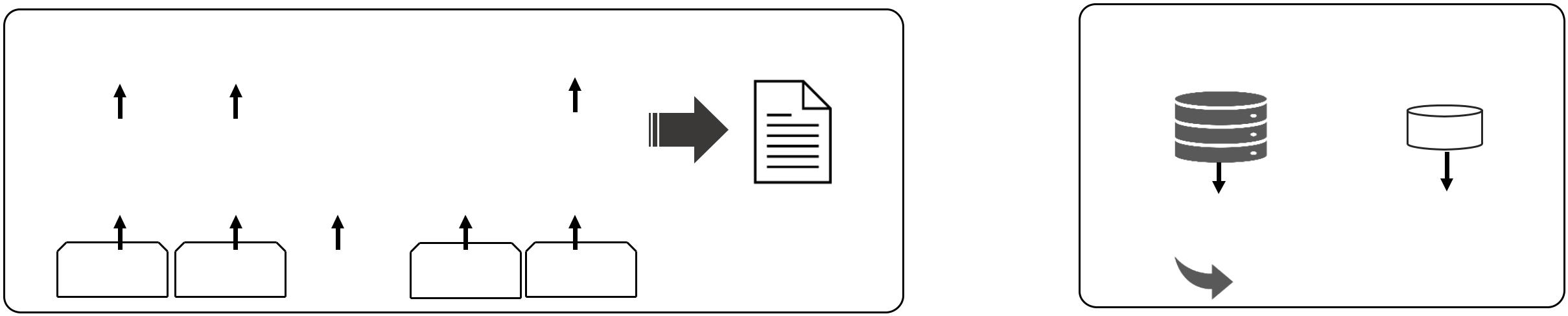
**ExecutablesDisassembledNormalized**

**(Tập văn bản)** **Chức năng** **Chức năng**

|  |  |  |
| --- | --- | --- |
| lớp hạ lưu | **tinh chỉnh** |  |
|  | **Mô hình BERT** |  |
|  |  |

* **Tiền xử lý để chuẩn bị đào tạo**③**Xây dựng một mô hình đặc biệt cho mã tương tự**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| huấn luyện viên trước | | | | | | | | | | | | | | | | | | | | | | | | | | |  | dự đoán | |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  | **Nhật ký** | | | | | | | | | | | | | | |  | Chức năng | | Mục tiêu |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | **…** | |  | nhúng | | Chức năng |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  | Mạng thần kinh chuyển tiếp nguồn cấp dữ liệu | | | | | | | | | | | | | | | | | | | | | | | |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  | BERT | | | | | | | | | | | | | |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  | **đào tạo trước** |  | Mô hình hạ lưu | |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  | [SOS]Số lần hiển thị 0 [MASK] Số lần hiển thị 2[EOS] | | | | | | | | | | | | | | | | | | | | | | | | |  |
|  |  | **Mô hình BERT** |  | Dự đoán: tương tự? | |  |
| **②Xây dựng một mô hình chung cho hội** | | | | | | | | | | | | | | | | | | | | | | | | | | |  | ④**Phát hiện sự giống nhau** | | |  |

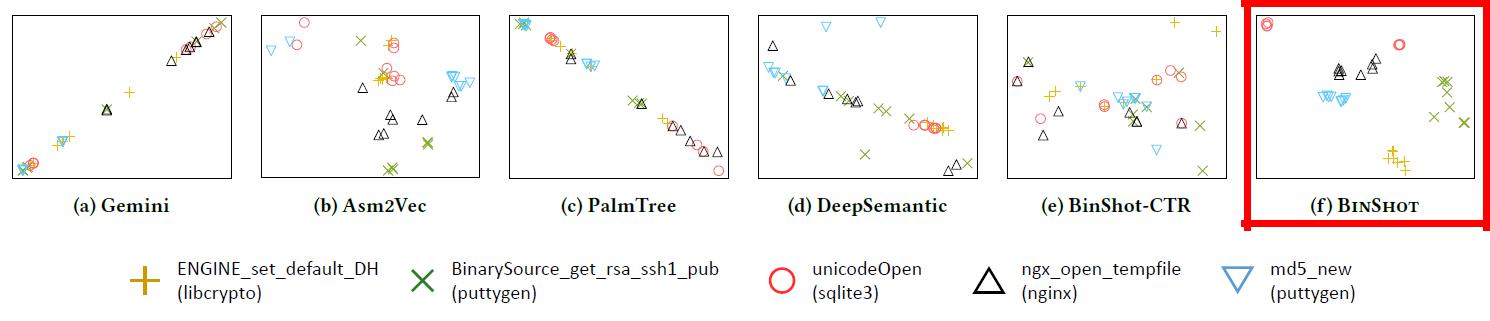
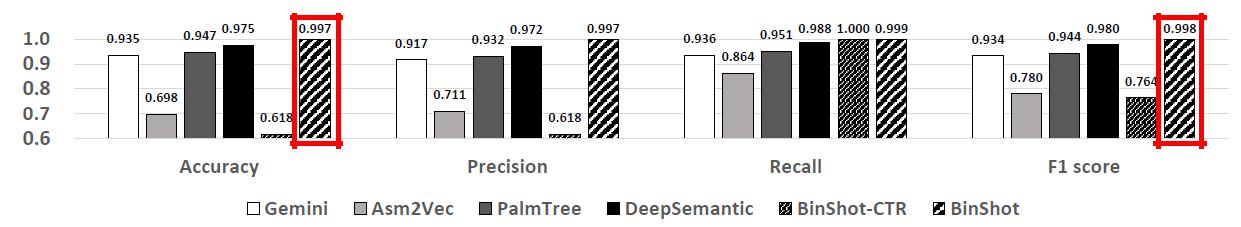


Thiết lập thử nghiệm

* tập dữ liệu
  + Được biên dịch với 2 trình biên dịch (gcc, clang) & 4 trình tối ưu hóa (O0-O3)
  + Tổng cộng 1.400 nhị phân
    - Tiện ích GNU – binutils, coreutils, diffutils, findutils
    - ĐẶC BIỆT2006, ĐẶC BIỆT2017
    - 11 Chương trình thực tế (BusyBox, Libgmp, …)
* Các mô hình cơ bản:
* Song Tử, Asm2vec, PalmTree, DeepSemantic
* BinShot-CTR, BinShot

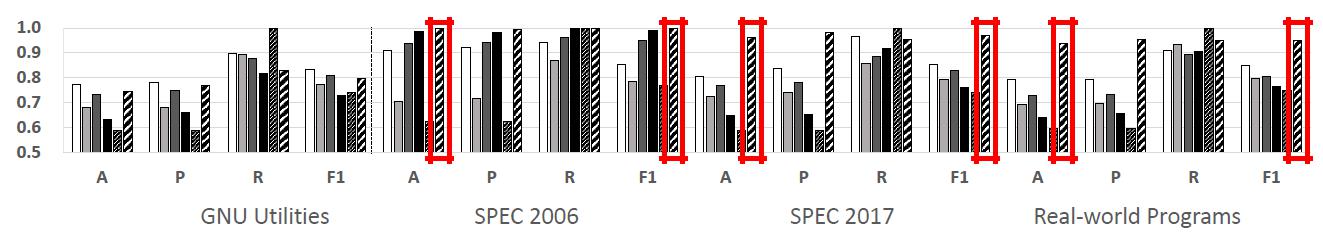
Đánh giá - Hiệu quả

* Đánh giá toàn bộ tập dữ liệu
* trực quan hóa t-SNE



Đánh giá - Khả năng chuyển giao

* Được đào tạo với SPEC 2006

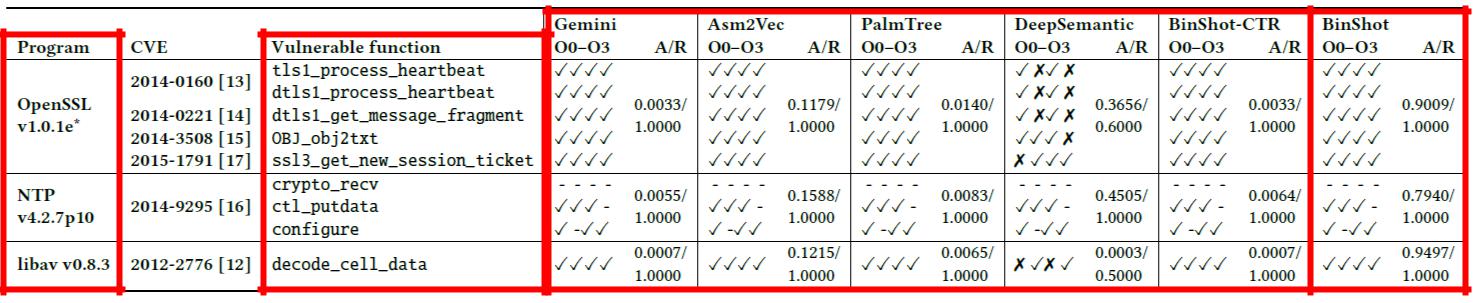


Đánh giá – Phát hiện chức năng dễ bị tổn thương

* Thiết lập kịch bản thực tế
  + Cơ sở dữ liệu chứa

nhúng chức năng dễ bị tổn thương

* + Truy vấn nhị phân bị tước
  + Mục tiêu: tìm một chức năng dễ bị tấn công từ một truy vấn nhị phân



Đánh giá – Hiệu quả thời gian chạy

* hiệu quả thời gian chạy
  + Exp1 - Từng cặp chức năng
  + Các cặp hàm Exp2 - 82300 (100 trong cơ sở dữ liệu, 823 trong nhị phân truy vấn) với công cụ dự đoán của chúng tôi

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Người mẫu | Song Tử | Asm2Vec | Cây cọ | Ngữ nghĩa sâu sắc | BinShot-CTR | BinShot |
|  |  |  |  |  |  |  |
| Exp1 (mili giây) | 0,10 | 81,94 | 1,33 | 1,34 | 1.30 | 1,32 |
|  |  |  |  |  |  |  |
| Exp2 (s) | 1.16 | 6.734,66 | 29.03 | 1,51 | 1,45 | 1,54 |
|  |  |  |  |  |  |  |

Thảo luận & Hạn chế



* Tên sai
* Chức năng nội tuyến
* Làm xáo trộn mã và các cấu trúc mã khác
* Hướng dẫn hiếm khi xuất hiện

Gói (lại

* Học khoảng cách có trọng số với entropy chéo nhị phân giúp cải thiện độ bền đối với các cặp hàm không nhìn thấy
* Tính ưu việt của BinShot
  + tính hiệu quả, tính thực tiễn (khả năng chuyển giao & thời gian chạy)
* Các mô hình khác nhưng mô hình của chúng tôi cho thấy hiệu suất kém trong một kịch bản thực tế
* Dự án mã nguồn mở: https://github.com/asw0316/binshot

Cảm ơn!